

Entropy and Stochastic Limits in LSH-Based Routing Systems

Many routing systems based on locality sensitive hashing (LSH) implicitly assume that increasing the number of routing bits yields finer specialization and greater representational capacity. However, increasing routing resolution eventually reveals irregularities in the routing distribution. These irregularities appear to arise from two fundamentally distinct mechanisms.

The Microscope Analogy

Increasing the number of routing bits may be viewed as progressively increasing the magnification of a microscope applied to the data distribution.

At low magnification, the routing space appears approximately uniform.

As magnification increases, two distinct limits are encountered.

1. The Entropy Limit

The **entropy limit** is reached when the routing resolution exceeds the intrinsic information content of the data.

Beyond this point, persistent irregularities appear in route occupancies because the data distribution itself is structured. Real-world datasets often occupy only a small subset of the ambient space, lying on low-dimensional manifolds, exhibiting sparsity, clustering, or strong feature correlations.

Consequently, many routing codes become intrinsically rare or impossible, even in the limit of infinite data.

Formally, if the routing distribution possesses entropy

$$H_D,$$

then routing resolutions satisfying

$$k > H_D$$

cannot be expected to exhibit uniform occupancy.

The entropy limit is therefore a property of the data distribution itself.

2. The Stochastic Limit

The **stochastic limit** arises from finite sample effects.

Even if the true routing distribution were perfectly uniform, a finite dataset containing N examples can only populate a finite number of routes. When the number of possible routes

$$2^k$$

becomes comparable to or exceeds N , occupancy fluctuations become dominated by sampling noise and many routes remain empty.

A rough criterion is

$$2^k \gtrsim N.$$

Beyond this point, route occupancy statistics become increasingly unreliable.

The stochastic limit is therefore a property of the dataset size rather than the underlying data distribution.

Implications for Atlas LSH Neural Networks

These limits may have important implications for Atlas LSH neural network design.

Large routing chunks, for example 20-bit selections, provide over one million possible routes. Unless both the entropy of the data and the dataset size are sufficiently large, many parameters may receive few updates or none at all.

This suggests several design principles.

- Prefer small routing chunks (for example 1-8 bits) to maintain high parameter utilization.
- Use hierarchical routing rather than large monolithic route selections.
- Measure empirical route entropy during training.
- Dynamically adapt routing resolution to observed occupancy statistics.
- Replicate heavily used routes and prune persistently unused routes.

An effective routing resolution

k_{eff}

may be estimated from the empirical routing entropy

$$H = - \sum_i p_i \log_2 p_i.$$

Increasing the nominal routing resolution beyond H is unlikely to provide significant additional specialization.

Implications for Conventional LSH Systems

The same considerations apply to traditional LSH lookup structures.

Increasing the hash length beyond the entropy limit causes many buckets to become permanently underutilized because the data occupy only a small subset of the possible code space.

Increasing the hash length beyond the stochastic limit causes bucket occupancies to become dominated by finite sampling effects, increasing memory requirements while reducing statistical reliability.

Consequently, the optimal hash length is not solely a property of the hash function, but rather depends jointly upon:

1. the intrinsic entropy of the data distribution, and
2. the available sample size.

This suggests that adaptive or data-dependent hash resolutions may outperform fixed-length hashing schemes in many practical settings.

In summary, LSH systems may be viewed as possessing both an **entropy limit**, determined by the intrinsic complexity of the data, and a **stochastic limit**, determined by finite sample size. Increasing routing resolution beyond either limit yields diminishing returns and may significantly reduce parameter utilization, lookup efficiency, and statistical reliability.